Program: B.Tech in Information Technology	**Academic Year: 2022**	**Duration: 3 hours**

Date: 10.01.2023
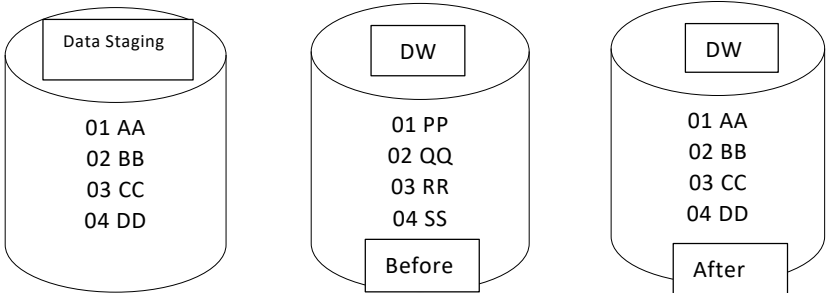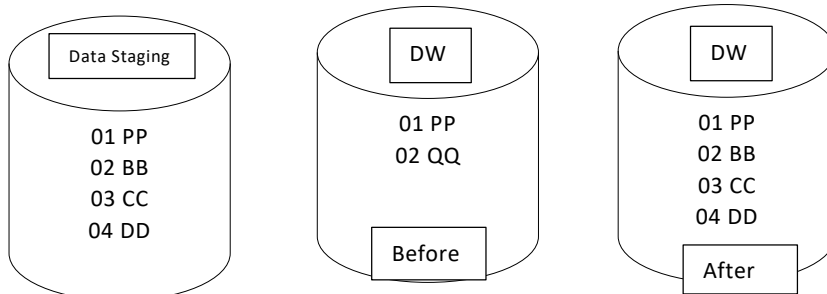**Time: 10:30 am to 01:30 pm**
Subject: Data Warehousing and Mining (Semester V)   **Marks: 75**

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.

(1) This question paper contains three pages.
(2) All Questions are Compulsory.
(3) All questions carry equal marks.
(4) Answer to each new question is to be started on a fresh page.
(5) Figures in the brackets on the right indicate full marks.
(6) Assume suitable data wherever required but justify it.
(7) Draw the neat, labelled diagrams, wherever necessary.

Question No.		Max. Marks
Q1 (a)	Why is data integration required in a data warehouse? Justify with the help of an example.	[05]
	OR	
	What is a factless fact table? Design a simple STAR schema with a factless fact table to track patients in a hospital by diagnostic procedures and time.	[05]
Q1 (b)	i. Describe slowly changing dimensions. List the types and explain any one type briefly with the help of an example.	[05]
	ii. Identify and justify the modes of applying data in data loading for the following scenarios:	[05]
	(A)	
	(B)	

Q2 (a)	What are hypercubes? How do they apply in an OLAP system?		[05]
	OR		
	Differentiate between MOLAP and ROLAP.		[05]
Q2 (b)	Consider a data warehouse for a hospital. Consider two measures i) Count ii) Charge where charge is the fee that the doctor charges a patient for a visit. For the above example create a cube and illustrate the following OLAP operations. 1) Rollup 2) Drill down 3) Slice 4) Dice 5) Pivot.		[10]
Q3 (a)	Consider a patient record TABLE 1 given below that contains attributes as Name, Gender, Fever, Cough and 4 tests, where Name is an object identifier. Find the distance between each pair of the 3 patients for asymmetric attributes. Also, mention the patients who are likely to have similar disease.		[05]

TABLE 1

NAME	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
JACK	M	Y	N	P	N	N	N
MARY	F	Y	N	P	N	P	N
JIM	M	Y	P	N	N	N	N

OR

	Consider the following data points: (13,15,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70) Find Q1, Q2, IQR, five number summary and Plot box plot.		[05]
Q3 (b)	Discuss different steps involved in data preprocessing.		[10]
Q4 (a)	i. Consider TABLE 2 given below and apply the Naïve Bayes Classification algorithm for "buys_computer" and classify the following tuple. X= {age=<=30", income= "medium", student= "yes" and credit_rating= "Fair"}		[05]
	ii. Apply K-means for the following dataset with two clusters. {1,2,6,7,8,10,15,17,20}		[05]
	OR		
	Apply ID3 on the following dataset (TABLE 2) from Electronics customer database and generate the decision tree.		[10]

TABLE 2

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Q4 (b)	Discuss the various categories of Clustering methods.	[05]

Q5 (a)	**Solve any one.**
	  i. Multilevel Association Rules and Multidimensional Association Rules
	  ii. Compare Clustering and Classification
		[05]

Q5 (b)	A dataset has 9 transactions mentioned in TABLE 3. Let minimum support = 2. Construct FP-Tree and Frequent patterns.	[10]

TABLE 3

TID	ITEMSETS
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C